

Název:

Elektronické uživatelské statistiky a jejich vývoj

Abstrakt:

Za elektronické informační zdroje a navazující poskytované služby platí knihovny vysoké finanční částky, proto je nutné velmi pečlivě sledovat a hodnotit jejich využití. Příspěvek se zabývá vývojem metod sběru elektronických uživatelských dat a jejich analýzou. Jsou shrnuty přínosy projektu COUNTER (Counting Online Usage of NeTworked Electronic Resources) a princip automatického sklizení dat pomocí protokolu SUSHI (Standardized Usage Statistics Harvesting Initiative), zmíněn je i relativně nový projekt PIRUS 2. Na závěr je nastíněn pravděpodobný budoucí vývoj v oblasti metod hodnocení využití EIZ.

Klíčová slova:

Elektronické uživatelské statistiky, vývoj, COUNTER, SUSHI, budoucí trendy, standardizace

Title:

Electronic Usage Statistics and Standards Development

Abstract:

Libraries spend huge amounts of money on electronic resources and related services; thus it is necessary to carefully measure and evaluate their usage. The article describes development of the methods for usage data collection and analysis. Described further is the principle of usage data harvesting via SUSHI (Standardized Usage Statistics Harvesting Initiative) and methods of COUNTER (Counting Online Usage of NeTworked Electronic Resources) statistics processing. The author gives a brief information about a relatively new project PIRUS 2. The presumable development of the methods of measuring usage closes the contribution.

Key words:

Electronic usage statistics, development, COUNTER, SUSHI, future trends, standardization

Elektronické uživatelské statistiky a jejich vývoj

Tištěné knihy nepředstavují velký problém v souvislosti se zjišťováním míry jejich využívanosti. Pokud jsou v knihovně evidovány výpůjčky, je snadné spočítat, kolikrát byla kniha zapůjčena. Data jistě nejsou zcela přesná, neboť nejsou evidovány prezenční výpůjčky z volného výběru, ale získaná data přesto mají značnou vypovídací hodnotu. Zjišťování míry využití časopisů je mnohem složitější. Kvantifikace jejich využití je stížena tím, že necirkulují mezi uživateli stejně jako knihy, jsou dostupné většinou jen prezenčně. Podíváme-li se na zjišťování využívanosti elektronických zdrojů je situace ještě složitější.

Než nakladatelé začali poskytovat uživatelská data, využívali knihovníci jiné zdroje informací, např. logy institucionálních webových stránek, statistiky dodávané poskytovateli A-Z seznamů a informace generované linkovacími servery. Všechna získaná data sice poskytla obraz o vzorcích chování uživatelů, ale rozhodně nenabízela přesné informace, které knihovníci potřebovali. Dokud neměla knihovna elektronický zdroj na svém serveru, byla závislá na datech obdržených od poskytovatele. Zpočátku neexistovala jednotnost v tom, co poskytovatelé změřili a často byly dodané informace nepoužitelné, nevypovídaly totiž nic o skutečném využití zdroje (např. do počtu navštívených stránek byly započítávány i stránky s nápovědou, menu apod.).

Na nátlak knihoven začali nakladatelé zkoušet měřit využití elektronických časopisů. Patricia Hults, Rensselaer Polytechnic Institute, USA, ve své studii [Hults, 2008, s. 31] uvádí, že uživatelé JSTOR byli v roce 1997 mezi prvními, kteří detailně popsali prvky, které chtějí mít obsažené ve statistických výstupech. Práce uživatelů JSTOR byla v roce 1998 rozšířena, když International Coalition of Library Consortia vydalo „Průvodce pro statistická měření využití webových indexových, abstraktových a plnotextových databází“ [ICOLC, 1998]. Přesto, že byla vydána tato dvě doporučení, informace od nakladatelů přicházely stále různorodé. Termíny jako „počet článků“, nebo „přihlášení do databáze“ zní jasně, ale způsob, jakým se počítají, může být rozdílný [Hults, 2008, s. 31]. Např. u počtu článků záleží na tom, jak byl článek rozčleněn, jestli se počítaly i odkazy na grafiku, jak se počítal tisk a stažení článku, jak bylo počítáno opětovné prohlížení článku v krátkém čase apod. Záleží také na tom, co článek obsahuje. Např. články zabývající se obchodem a ekonomikou často obsahují statistické grafy, které nejsou součástí psaného textu. Nekonzistentnost se objevuje i v počtech přihlášení a vyhledávání v databázi (sessions and searches). A můžeme se ptát dál: Jak byla počítána opětovná přihlášení jedním uživatelem v krátkém čase? Jak se počítalo vyhledávání ve více databázích najednou – jako jedno, nebo dvě? S využíváním federativního vyhledávání se měření ještě více komplikuje.

Dalším problémem bylo, že způsob měření využívanosti zdroje u jednoho nakladatele se mohl během nějaké doby změnit. Knihovny pak mohly mít od jednoho nakladatele nekonzistentní data. Jestliže nakladatel opravdu změnil způsob výpočtu využití zdroje, knihovna musela žádat o konverzi dat do nového formátu, pokud to ovšem vůbec bylo možné. Jinou odlišností byla retrospektiva dat poskytovaných nakladateli a frekvence sklizení dat. Způsob poskytování statistických dat se také nakladatel od nakladatele lišilo. Někdy měli uživatelé možnost vytvářet statistiky samostatně, jindy bylo získání dat omezené na zasílání mailem. Formáty reportů se také mohly velmi lišit.

Projekt COUNTER – zlom ve využití elektronických statistických dat

V roce 2000 nabízela uživatelské statistiky méně než polovina nakladatelů elektronických časopisů [Luther, 2000, s. 1]. V roce 2000 a 2001 bylo vydáno několik podstatných studií zabývajících se otázkou standardizace statistik pro elektronické zdroje. V roce 2000 to byl „White Paper on Electronic Journals Usage Statistics“ a ARL (Association of Research Libraries) iniciovala projekt E-metrics¹. V

roce 2001 vydaly své dokumenty ke standardizaci statistik americká National Commission on Libraries and Information Science a NISO (revize standardu Z39.7). Tyto rané práce jsou základem projektu COUNTER², který vedla skupina PALS (Publishers and Librarian Solutions). Pracovní skupina pro uživatelské statistiky v rámci PALS vytvořila základní dokument, tzv. „code of practice“, ve kterém jednoznačně určila a definovala datové prvky, doporučila datové formáty a způsoby dodávání dat. Tato práce se vyvinula v projekt COUNTER, který byl oficiálně vyhlášen v roce 2002. Úspěch projektu je z velké části zásluhou spolupráce knihoven, nakladatelů a dalších profesních společností. Dokument „code of practice“ vyšel poprvé v lednu 2003. V březnu 2011 už k projektu COUNTER byli přihlášení všichni velcí a velký počet menších poskytovatelů/nakladatelů elektronických časopisů. V tabulce na webových stránkách projektu je uveden seznam, ve kterém figuruje 131 subjektů³, v roce 2006 to bylo o polovinu méně - 51 [HULTS, 2008, s. 32].

1. vydání „code of practice“ pokrývalo pouze časopisy a databáze. Dokument definoval základní pojmy „vyhledávání“ a „stažení článku“, co má report obsahovat, v jakém formátu má být přístupný a jak často má být generován. Jakmile začali nakladatelé dodržovat určené standardy, bylo možné porovnávat a hodnotit využití napříč uživatelskými platformami.

1. vydání „code of practice“ pro elektronické knihy a referenční práce vyšlo v roce 2006. Stejně jako u časopisů a databází byly definovány datové prvky. Statistiky pro elektronické knihy a encyklopedie jsou ještě o něco problematičtější než statistiky pro časopisy a databáze. Není totiž tak snadné definovat základní jednotku využití knihy jako u časopiseckého článku. Paralelně s časopiseckým a databázovým reportem byl definován i obsah reportu pro knihy a encyklopedie.

2. vydání „code of practice“ pro časopisy a databáze, vydané v roce 2006, nerozšířilo počet datových prvků. Byla dána přednost redefinování obsahu reportů a důraz byl kladen především na přesné určení kroků, které museli provést nakladatelé, aby byly dodrženy standardy. Aby se standardy staly šířeji využívané, bylo nutné definovat metody pro ověření správnosti měření využívanosti elektronických časopisů. Bylo nutné zjistit, jestli nakladatel skutečně měří, co COUNTER specifikuje jako standard. Metody auditu zahrnují velmi specifické kroky. Nejdříve nakladatel upozorní zástupce COUNTER na to, že má zájem o zapojení do projektu a je ochoten poskytovat statistiky podle standardu. Pak následuje testování reportů knihovnamí, které byly k tomuto účelu vybrány. V současné době jsou to následující knihovny: Cranfield University (UK), University of California (USA), University of Leicester (UK), University of Liverpool (UK), University of Wales at Bangor (UK), Yale University, (USA) [COUNTER, 2011]. Certifikát obdrží nakladatel až tehdy, pokud jsou testovací knihovny spokojeny. Poté, co je nakladatel schválen knihovnamí, má 18 měsíců na to, aby dokončil audit. Musí např. provést určitý počet vyhledávání a stažení článků, sledovat časové intervaly mezi vyhledáváním a stahováním dat, získaná data pak porovnat. Po vstupním auditu následují audity roční, na základě kterých se nakladateli ponechává, nebo odebírá certifikát.

3. vydání „code of practice“ vyšlo v dubnu 2008. Nakladatelům jsou kladeny podmínky, aby poskytovali SUSHI služby, aby byla dodávána data o využívání zdrojů přes služby federativního vyhledávání a nově jsou také definovány podmínky pro dodávání dat pro konsorcia [GROGG; FLEMING-MAY, 2010, s.12]. Webové stránky projektu COUNTER obsahují kompletní tři vydání a detailní informace o podmínkách platných pro splnění auditu pro nakladatele. COUNTER společně s United Kingdom Serials Group (UKSG) realizovaly výzkum o tom, jak mohou uživatelské statistiky přispět k hledání nových cest měření kvality časopisů – Journal Usage Factor. Závěrečná zpráva z výzkumu byla vydána v roce 2007⁴.

Přestože COUNTER ušel ke standardizaci statistik pěkný kus cesty, stále tu jsou ještě jistá omezení

týkající se především způsobů zpřístupnění elektronických zdrojů. Např. jestliže univerzitní knihovna předplácí stejný titul jako jednotlivec na univerzitě (vědec, pedagog), není možné oddělit uživatelské statistiky na dva účty, data se generují společně do jednoho reportu. Dalším příkladem mohou být databáze, ve kterých jsou zpřístupněny předplacené tituly společně s dočasně volně přístupnými tituly (trial), nebo open accessovými tituly. Oddělit statistická data předplacených a nepředplacených časopisů z jedné databáze je příliš komplikované [HULTS, 2008, s.33].

Autoři Davis a Price [DAVIS; PRICE, 2006] uvádějí, že rozdíl využívanosti jednoho zdroje může způsobit vzhled platformy a její funkcionalita. Zkoumali poměr stažených článků ve formátu HTML a PDF v souvislosti s platformami. Autoři dospěli k zajímavým výsledkům. Např. byl zkoumán jeden časopisecký titul přístupný prostřednictvím dvou rozdílných platforem a potvrdil se předpoklad, že platforma ovlivňuje poměr využití formátů. Některé platformy totiž upřednostňují formát HTML, až stránka s plným textem ve formátu HTML odkazuje na plný text ve formátu PDF. Některé platformy naopak umožňují stažení článku ve formátu PDF již na úrovni citace, nebo abstraktu, čímž je redukován počet stažení článků ve formátu HTML, následkem toho i celkový počet stažení.

PIRUS 2

Další iniciativou v oboru uživatelských statistik je projekt PIRUS 2⁵, který staví na standardech COUNTER a výsledcích původního projektu PIRUS (Publisher and Institutional Repository Usage Statistics). Tento projekt demonstroval, že je možné unifikovat uživatelské statistiky pro jednotlivé články navzdory tomu, že jsou roztroušeny v organizačně a technicky rozmanitých prostředích. PIRUS 2 má za cíl vytvořit sadu standardů, protokolů a procesů, které budou umožňovat nakladatelům, repozitářům a dalším organizacím vytvářet autoritativní, spolehlivé statistiky využití jednotlivých článků a dalších jednotek, které zpřístupňují. Digitalizace zásadně změnila způsoby práce vědců. Zájem se obrací více na menší části tradičních médií, což nutí knihovnické a nakladatelské komunity hledat nové cesty pro získávání a sdílení uživatelských statistik na úrovni článků/kapitol. PIRUS 1 byl veden členy COUNTER a skončil v lednu 2009. PIRUS 2 probíhal od října 2009 do února 2011 a byl veden Mimas a Cranfield University. V projektovém týmu byli i zástupci COUNTER, CrossRef a Oxford University Press. CrossRef je důležitý člen týmu, protože reprezentuje rozsáhlou komunitu nakladatelů. Je to největší registrační agentura pro DOIs (Digital Object Identifiers) a nadále bude mít vedoucí postavení ve vývoji technologie, která umožňuje linkování všech typů vědeckých prací přes DOI a OpenURL [GROGG; FLEMING-MAY, 2010, s.12].

SUSHI – automatické sklizení dat

Ve stejné době, kdy byly vyvinuty standardy pro uživatelské statistiky, paralelně probíhal vývoj standardů pro správu EIZ. ERMI (Electronic Resource Management Initiativ) v rámci Digital Library Federation ⁶ začala s definicí datových prvků nezbytných pro efektivní správu EIZ. ERMI následovala ERMI 2, která navíc obsahovala požadavky na uživatelská statistická data, jejich získávání a analýzu [DLF ERMI, 2006]. V létě roku 2005 vznikla malá pracovní skupina, která se zabývala možností přenosu statistických dat od nakladatele do systému ERM. Skupina se skládala ze tří knihovníků (Ivy Anderson, California Digital Library; Adam Chandler, Cornell University Library; Tim Jewell, University of Washington Libraries) a zástupců čtyř komerčních společností (Ted Fons, Innovative Interfaces, Inc.; Bill Hoffman, Swets Information Services; Ted Koppel, Ex Libris; Oliver Pesch, EBSCO Information Systems) [CHANDLER; JEWELL, 2006, s. 1]. Z práce této skupiny vznikl projekt SUSHI (Standardized Usage Statistics Harvesting Initiative), který vzala pod svá křídla NISO. Původní pracovní skupina se rozrostla o několik nakladatelů a producentů automatických knihovnických systémů. SUSHI počala vyvíjet protokol založený na standardech COUNTER, který by

umožnil automatické sklizení dat od různých nakladatelů s následnou distribucí sklizených dat do systému ERM. V říjnu 2007 byly výsledky práce skupiny shrnuty a publikovány v dokumentu pod názvem ANSI/NISO Z39.93-2007 - The Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol.

Iniciativa SUSHI vznikla jako odpověď na spoustu času a práce knihovníků, kterou věnují ručnímu sběru uživatelských dat. Základní koncept SUSHI je jednoduchý. Na serveru knihovny je SUSHI klient, který je často integrován se systémem ERM, na straně poskytovatele je SUSHI server. Systém ERM by měl být schopný automaticky zasílat požadavky, sklízet a integrovat statistiky poskytované nakladateli, bez zásahu lidské ruky. Pro výměnu dat bylo vybráno schéma XML. Ideálně systém ERM zformuluje požadavek, uloží všechny potřebné informace do souboru XML a soubor odešle na SUSHI server, který statistiky poskytuje. SUSHI server by měl být schopen informaci přijmout, vygenerovat požadovaná data ve formátu COUNTER, vložit je do souboru XML a poslat zpět systému ERM, který požadavek původně vyslal. Pokud se vyskytne nějaká chyba, měla by být systému ERM zaslána chybová zpráva. Systém ERM by poté měl umět zprávu rozbalit a uložit ji do své databáze. Nakonec by měl systém ERM generovat náležité reporty. Protokol SUSHI zajišťuje automatický přenos informací ze systému ERM k nakladateli a zpět. [CHANDLER; JEWELL, březen 2006, s. 68].

Pravděpodobný budoucí vývoj měření využívanosti zdrojů

Knihovníci už dávno nejsou jen „sběrateli“ informací. Prudký rozvoj institucionálních repozitářů způsobil, že knihovnici se stali také distributory informací. Plnění a údržba repozitářů je nákladnou záležitostí a knihovny chtějí zpětnou vazbu, proto se snaží získat nějaká data o jejich využití. Na světě jsou nyní stovky institucionálních repozitářů, ale zatím neexistuje standard, který by jejich využití uměl změřit. Členům projektu COUNTER bylo už v roce 2007 jasné, že sbírání statistických dat na úrovni časopisů nebude brzy stačit. Pracovní skupina pro uživatelské statistiky projektu COUNTER proto začala pracovat na vývoji rovnice pro výpočet využití plných textů článků tzv. Journal Usage Factor (JUF). V porovnání např. s Journal Impact Factorem (JIF) bude JUF mnohem aktuálnější, protože data jsou sbírána hned po publikování článku, JIF je založen na výpočtech citačních aktivit a na podklady se čeká dva roky [KING, 2009, s.5]

Úkolem první fáze projektu COUNTER JUF, financovaného UKSG (U.K. Serials Group), bylo ověřit proveditelnost vývoje a implementace JUF. Ukázalo se, že koncept JUF je nejen smysluplný, ale také výrazně podporovaný nakladateli, knihovníky a výzkumnými komunitami. Na základě pozitivních výsledků první fáze projektu se rozhodla UKSG společně s dalšími institucemi financovat projekt dále. Cílem druhé fáze projektu bylo posoudit zda bude JUF solidní, implementovatelný a finančně efektivní nástroj pro zjištění relativního postavení a hodnoty časopisu založeného na testování samostatných prvků (článků) s využitím pravdivých uživatelských dat poskytnutých různými nakladateli [SHEPHERD, 2011]. Stanovení výpočtu JUF je ve finální fázi. Před tím, než bude JUF formálně uznán jako nový standard, musí však ještě proběhnout další analýzy testovacích dat.

Už několik let existuje řada dalších projektů, jejichž cílem je vyvinout vhodnou metodu pro získávání spolehlivých dat o využívání samostatných článků. Jedním z takových projektů je již výše zmiňovaný projekt PIRUS. V budoucnu bude obzvláště žádoucí, aby nakladatelé a tvůrci repozitářů přijali stejný standard pro měření využití samostatných jednotek. Jedině tak bude možné získat důvěryhodná data o využití informačních zdrojů.

Podle Jana Lewise z East Carolina University budou za pět let knihovny trvale osvobozeny od práce se sbíráním statistických dat. Díky plné automatizaci se uvolní ruce knihovníkům a budou moci věnovat čas jiným činnostem souvisejícím s akvizicí a tvorbou sbírek. Současný pracný způsob získávání

statistických dat není uspokojivý, ale je daleko snadnější a standardizovanější, než tomu bylo před pěti lety [KING, 2009, s.5].

Uživatelské statistiky jsou standardizovanější a snadno se získávají, roste počet institucí, které je sbírají, analyzují a budou jim přikládat větší váhu v evaluačním procesu. Předpokládá se, že všichni poskytovatelé EIZ budou zpřístupňovat standardizované statistiky. Na trhu se objevují nové produkty, které umožňují sklizení a analýzu uživatelských statistik jako např. ScholarlyStats od EBSCO, nebo Journal Use Reports od Thomson Reuters, ale jen čas ukáže, zda to i nadále budou životaschopné produkty, neboť některé knihovny už samy umí automaticky sklízet a analyzovat svá vlastní data.

Literatura:

BEALS, Nancy; LESHER, Marcella. 2010. Managing electronic resource statistics. *The Serials Librarian*. 2010, vol. 58, s. 219-223. ISSN 1541-1095.

COUNTER : Counting Online Usage of NeTworked Electronic Resources. 2011, [cit. 2011-03-13]. Dostupné z WWW: <<http://www.projectcounter.org>>

COUNTER : Counting Online Usage of NeTworked Electronic Resources. 2011. Register of Vendors. březen 2011, [cit. 2011-03-13]. Dostupné z WWW: <<http://www.projectcounter.org/compliantvendors.html>>

DAVIS, P.M.; PRICE, J.S. 2006. eJournal interface can influence usage statistics: Implications for libraries, publisher, and Project COUNTER. *Journal of the American Society for Information Science and Technology*, vol. 57, issue 9, s. 1243-1248.

DLF Electronic Resource Management Initiative, 2006. Phase II., [cit. 2011-03-13]. Dostupné z WWW: <<http://old.diglib.org/standards/dlf-erm05.htm>>.

FLEMING-MAY, Rachel A.; GROGG, Jill E. 2010. Standards, tools, and other products. *Library Technology Reports*. Sep 2010, vol. 46 Issue 6, s. 11-16. ISSN 0024-2586.

HULTS, Patricia. 2008. *Electronic Resource Management in Libraries : Research and Practice*. New York : Scott Breivold, 2008. Chapter II. : Electronic Usage Statistics, s. 29-46.

CHANDLER, A.; JEWELL, T. 2006. Standards – libraries, data providers, and SUSHI : The Standardized Usage Statistics Harvesting Initiative. *Against the Grain*. 2006, vol. 18, issue 2, s. 1-2. ISSN 1043-2094

CHANDLER, A.; JEWELL, T. březen 2006. The standardized usage statistics harvesting initiative (SUSHI). *Serials*. 2006, vol. 19, issue 1, s. 68-69.

ICOLC : International Coalition of Library Consortia (ICOLC). November 1998. Guidelines for statistical measures of usage of web-based indexed, abstracted, and full text resources. March 2011, [cit. 2011-03-16]. Dostupný z WWW: <<http://www.library.yale.edu/consortia/webstats.html>>.

LUTHER, Judy. 2000. White paper on electronic journal usage statistics. Washington, D.C.: Council on Library and Information Resources. [cit. 2011-03-23]. Dostupný z WWW:

<<http://www.clir.org/PUBS/reports/pub94/contents.html>>

KING, Douglas. 2009. E-opinions from the field : what is the next trend in usage statistics in the libraries? *Journal of Electronic Resources Librarianship*. 2009, vol. 21, s. 4–14. ISSN 1941-126X.

McQUILLAN, Bob; FATTIG, Karl Maria; KEMP, Rebecca; STAMISON, Christine; ENGLAND, Deberah. Electronic resource management system integration strategies: opportunity, challenge or promise?. *Serials Librarian*, Jan-Jun 2010, vol. 58, issue 1-4, s. 106-116. ISSN 1541-1095.

SHEPHERD, Peter T. 2011. Journal Usage Factor – a promising new metric. *Serials*. 2011, vol. 24, issue 1, s. 64-68. Dostupný také komerčně z WWW (doi): <<http://dx.doi.org/10.1629/2464>>. ISSN 0953-0460.

Poznámky:

1. <http://www.arl.org/stats/initiatives/emetrics/index.shtml>
2. Counting Online Usage of Networked Electronic Resource - <http://www.projectcounter.org/>
3. http://www.projectcounter.org/R1/R1_CV_Mar2011.pdf
4. <http://www.uksg.org/usagefactors>
5. <http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php>
6. <http://old.diglib.org/standards/dlf-erm02.htm>